Federal Aviation Administration

DOT/FAA/AM-05/15
Office of Aerospace Medicine
Washington, DC 20591

Pilot Willingness to Take Off Into Marginal Weather, Part II: Antecedent Overfitting With Forward Stepwise Logistic Regression

William Knecht
Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

August 2005

Final Report

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-05/15		

4. Title and Subtitle	5. Report Date
Pilot Willingness to Take Off Into Marginal Weather, Part II: Antecedent Overfitting with Forward Stepwise Logistic Regression	August 2005
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Knecht WR	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125	11. Contract or Grant No.
12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	14. Sponsoring Agency Code

15. Supplemental Notes

Work was accomplished under approved task AM-HRR-522-04.

16. Abstract

Adverse weather is the leading cause of fatalities in general aviation (GA). In prior research, influences of ground visibility, cloud ceiling height, financial incentive, and personality were tested on 60 GA pilots' willingness to take off into simulated adverse weather. Results suggested that pilots did not see "weather" as a monolithic cognitive construct but, rather, as an interaction between its separate factors. However, methodological issues arose during the use of logistic regression in modeling the effect of 60+ candidate predictors on the outcome variable of takeoff into adverse weather. It was found quite possible to obtain false "significance" for models comprised merely of random numbers, even when the number of model predictors was limited to a conventional 1/10. Therefore, Monte Carlo simulations were used to derive unbiased estimates of model significance and R^2 values. Research in correction for this case/candidate predictor ratio effect is relatively new and noteworthy, particularly in the social sciences. It was given the name "antecedent overfitting" to contrast with the more commonly known "postcedent" type, which is based on a small case/model predictor ratio.

17. Key Words	18. Distribution Statement
Statistics, Methodology, Overfitting, Case/Predictor Ratio, Regression	Document is available to the public through the Defense Technical Information Center, Ft. Belvior, VA 22060; and the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	17	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

ACKNOWLEDGMENTS

The author gratefully acknowledges the direction and comments of Barbara Tabachnick and, most particularly, the help of Robert Stine, Department of Statistics, Wharton School, University of Pennsylvania, who reviewed this manuscript.

Pilot Willingness to Take Off Into Marginal Weather, Part II: Antecedent Overfitting With Forward Stepwise Logistic Regression

INTRODUCTION

Part I of this report was entitled *The Influence of Visibility, Cloud Ceiling, Financial Incentive, and Personality Factors on General Aviation Pilots' Willingness to Take Off Into Marginal Weather*. In Part I, we reviewed data and made preliminary conclusions from a study of VFR takeoff into marginal weather conditions. At that time, we made reference to a number of statistical issues, some of which were to be deferred to a Part II report. This is that second report. In it will be addressed both the relevant statistical concerns that were uncovered plus the effect these had on the interpretation of the Part I results.

A problem naturally comes when some experimental situation forces us to deviate from routine procedure. Specific to the situation here, we had examined a large number of predictors with logistic regression (originally 83, finally reduced to about 60). It is standard statistical practice to limit the number of predictors included *within* any given regression model, usually to a ratio of about one predictor per 3-10 cases examined (Tabachnick & Fidell, 2000; R.A. Stine, personal communication, January 26, 2004). Otherwise, the data may be *overfitted*.

In its usual context, overfitting refers to the ability of a relatively large predictor/case ratio to mimic an arbitrary mathematical function. This phenomenon has long been known; in fact, it finds its origin in such useful mathematical fundamentals as the Taylor series and Fourier series (Kreyszig, 1972, p. 574, Taylor series). For example, the seemingly complex waveform in Figure 1 (left) can actually be broken down into the sum of a small number of discrete component sine waves, each with its own amplitude, period, and phase (right). This is a creative use of this kind of curve-fitting principle, whereby we take something complicated and explain it in simpler terms.

But there is a sinister side to the same idea that applies directly to regression analysis. If we try to include too many predictors into the standard sigmoid (S-shaped) logistic regression model below,

$$p_{event} = \frac{1}{1+e^{-(\beta_0+\beta_1 P_1+\beta_2 P_2 ...+\beta_n P_n)}} \quad (1)$$

we can actually overfit the model by juggling the β (beta) coefficients in the exponent, $\beta_0 + \beta_1 P_1 ...$, until we arrive at a prediction function that superficially seems to fit our data fairly well. However, that fit can owe more to this general ability to fit anything with enough terms than it does to our actual ability to find a small number of valid, reliable factors truly modeling real, underlying processes.

Overfitting is usually considered worst when it inflates Type I error (false statistical significance when none truly exists in the population). Ideally, Type I error should only reflect sampling error—pure variation due to subject-related factors. In fact, we *expect* Type I errors about 5% of the time with normally distributed random numbers when the statistical significance level is set at $\alpha = .05$—because that is precisely how "$\alpha = .05$" is defined in the first place.

But Type I error can also be an unwanted side effect of ill-considered experimental design or statistical method. And this is where this issue of overfitting relates to our Part I experiment. At some point during the analysis of those data, an intuition arrived. If forward stepwise regression were performed on *n* cases (pilots), starting with a large set of *p* candidate predictors, could overfitting occur even though only a small number *k* of predictors were allowed into any given model? Could this happen even if the predictor/case ratio (*k*/*n*) were maintained strictly at, say, 1/10 per model, as one common rule of thumb dictates? Was it a problem that we had 60 candidate predictors, even if no model were allowed more than one predictor per ten pilots?

In other words, could there be *two* kinds of overfitting, only one of which is normally mentioned in statistical texts written for the social sciences? This issue had more than passing practical significance—if it were not settled, it could call into question all the conclusions in the Part I study.

No textbook any of us had seen mentioned this specific problem. We had seen overfitting discussed only relative to the number *k* of predictors in the *model*, not the number *p* of *available* predictors. In Part I, we used Statistical Packages for the Social Sciences V11.5 to do the logistic regression (Norušis, 1999; SPSS, 2003). This contained no adjustment for *p*, nor was any word of this issue found in the software documentation or on the SPSS corporate Web site.

METHOD

A Quick Random Number Simulation

To test our suspicions with controlled data, a normal-random data set was generated in Excel 2000 (Microsoft, 1999). This set emulated data from 30 pilots. Each pseudo-"pilot" had 60 random "predictor scores" generated by Excel's normal distribution pseudo-random function. A pseudo-random function is a mathematical equation that generates a distribution of numbers which, over the course of many iterations, behaves like a sample from a truly random distribution (Press, Flannery, Teukolsky, & Vetterling, 1988, ch. 14). In this case, each pseudo-predictor score was based on a mean of 5, and standard deviation (SD) of 1. The exact choice of mean and SD should not be critical, since logistic regression adjusts the relative contribution of each predictor by multiplying it by its own β coefficient. Below is an abbreviated example of what this random data set looked like. The random scores themselves are highlighted in gray.

The structure of this data set closely paralleled the Part I technical report, which did compare two sets of 30 pilots, each having about 60 predictor scores per pilot. Those real predictor scores had been measurements taken on various environmental conditions, pilot demographics, and responses on a number of psychological personality tests.

Our new, randomly generated data were next run through SPSS forward stepwise, likelihood-ratio logistic regression, using *Takeoff* as the dependent variable, the same way as was done for the Part I Low Financial Incentive experimental group (n=30). The dichotomous dependent variable (DV) *Takeoff* was coded as 0 for "No" and 1 for "Yes." The success ratio (pilots taking off / total pilots) was set at (9/30) = .30, just as the actual Part I results had been. SPSS then proceeded to select three of the 60 random pseudo-predictors as "best," and calculated a factor-weighted prediction score, namely

$$p_{event} = \frac{1}{1+e^{-(1.146P_2+1.725P_4-3.084P_{40})}} \qquad (2)$$

(See Figure 2 of above graphed calculation.)

Each of the 30 "pilots" above was represented by its own number, 1-30, on the x-axis. Note how each had three pseudo-predictor score values, P2, P4, and P40, which SPSS logistic regression selected as best from the total set of 60. The actual pilot takeoff score (heavy dashed line, value 0-1), was a step function with "Take-off" represented by 1, and "No takeoff" as 0. Finally, notice the prediction score (solid "Prediction Eq." line, also 0-1, the result of Equation 2). This ran quite close to the actual takeoff score, implying a very good fit of predicted takeoff to actual takeoff.

The point of this whole exercise was to test quickly if a group of random numbers could predict a high percentage of takeoffs. This example showed that it could. Predictivity[1] was (27/30) = 90%. Yet, this was completely due to SPSS acting on nothing but noise. Look at the raw scores themselves, P2, P4, and P40. There was no particular pattern to, or correlation between, these three predictors. The only pattern was in the weighted sum ($\beta_2 P_2 + \beta_4 P_4 + \beta_{40} P_{40}$) after it was run through the SPSS modeling algorithm. This did not imply that anything was wrong with SPSS logistic regression. What it implied was the presence of some deeper statistical phenomenon at work, one undocumented in the textbooks we had read.

This was initially unnerving, since it seemed to call into question many of our Part I conclusions. How it could happen was not that surprising, though, as we began to consider the situation in detail. Theoretically, there were (60•59•58)/(3•2•1) = 34,220 possible three-predictor models to choose from.[2] And, even though stepwise regression does not examine all possible combinations, it still does "capitalize on chance variation" (Derksen & Keselman, 1992). It starts by first finding the single best predictor and then adds others, according to their relative improvement to the model. It is hill-climbing in predictivity space.[3] And, even though hill-climbing does not guarantee getting to the absolute highest possible predictivity, it normally gets to one of the higher peaks. And here this was happening with random numbers. It all goes to show that even rare events may become quite likely when we roll the dice often enough (have too many predictors) or reach into the jar and feel for the biggest marbles (use stepwise regression). This is not to say these techniques should be strictly forbidden, it simply says we need to exercise caution.

So, to summarize, this kind of overfitting was not the same as that discussed in the average social sciences statistics textbook. Instead, the problem centered around the large number of predictors available before we started modeling. For this reason, it could be called *antecedent overfitting*, because it derives from a condition existing prior to the analysis. The more common kind of

overfitting—having too many predictors inside a given model—could then be more aptly called *postcedent overfitting*, since that has to do with events occurring *after* the number of candidate predictors is already established. In antecedent overfitting, the number of candidate predictors p is too large. In postcedent overfitting, the number of predictors k included in the final model is too large.

Literature Search

Once it became obvious that this problem was a legitimate challenge to the Part I analysis, the next step was to consult a nationally known statistician (Tabachnick, personal communication, May 15, 2003). She confirmed the suspicion that, if known at all, the topic was not common in the social sciences. An extensive Internet search finally led to a 1998 unpublished draft of a paper by Foster and Stine that directly referenced this problem in standard linear regression (p. 2): "This tendency of stepwise regression to overfit grows with the number of available predictors, particularly once $p > n$" (n being the number of cases). This article allowed back referencing to other key studies, Rencher and Pun (1980), and Kendall and Stuart (1961, ch. 27), all having to do with conventional linear regression.

So, it appears that this problem has been known in linear regression for at least 40 years. However, it has been largely ignored outside the professional statistics community, nor has the extension to logistic regression yet been published (R.A. Stine, personal communication, July 27, 2003).

Monte Carlo Simulations

A single look is insufficient to reliably explore a phenomenon. What would be sufficient here would be either a) closed-form solutions for the maximum likelihood estimators (MLE) and confidence intervals (CI) of both predictivity and R^2, and/or b) Monte Carlo simulations to arrive at the same estimates.[4] A closed-form solution is a single, globally optimal or correct solution that can be expressed as a solvable mathematical equation (e.g. $y = 3x + 4$). To a statistician, closed-form solutions are always the ideal. However, for a number of reasons, it is sometimes impossible to find closed-form solutions. In that case, the standard procedure is to use numerical methods—for example computer algorithms using pseudo-random numbers as input to repeat some statistical computation hundreds or thousands of times, until the outcomes achieve some desired level of statistical stability/reliability. Monte Carlo simulations are such numerical methods, used when it is impossible to find a closed-form solution. They are also used to cross-check the validity and accuracy of closed-forms.

In our particular case, logistic regression has no closed-form solution. Instead, results are calculated using a set of equations (SPSS, 2003) run through an algorithm (i.e., a rule-based set of instructions). Most of the time this algorithm produces valid results, but there can be times when it fails (R.A. Stine, personal communication, January 26, 2004; Tabachnick & Fidell, 2000, p. 522). Strangely enough, this happens whenever a single predictor has 100% predictivity and can successfully classify all DV outcomes. This causes the algorithm's Newton-Raphson estimation of model parameters to go wildly out of control and head off toward zero or infinity (Press, et al., 1988, ch. 9.4). To guarantee termination, the SPSS algorithm simply halts after a certain number of iterations, but the resulting model parameters are nonsensical. A second way the logistic regression algorithm can halt is bootstrap failure.[5] In that case, the algorithm cannot get beyond the very first step, because no predictor meets even the minimum criterion for inclusion (SPSS calls this the "PIN"). Predictors are included in forward stepwise regression because they improve model performance to some prespecified degree. Calculation stops when having more predictors fails to bring the specified degree of improvement.

Lacking closed forms for model parameter estimates, and lacking the ability to derive such estimates ourselves, the present investigation was limited to Monte Carlo simulations. This would at least allow us to correctly estimate the following critical information for our Part I Low and High Financial Incentive models, both with and without a constant:

1. Mean Predictivity	A ratio μ_p: (mu, cases successfully predicted / total cases)
2. Standard deviation of predictivity	σ_p (sigma)
3. Mean Nagelkerke R^2	A ratio μ_{R2}: (variance explained / total explainable variance)
4. Standard deviation of Nag. R^2	σ_{R2}
5. .95 confidence intervals	Predictivity and R^2 necessary for a model to arguably exceed chance

However, keep in mind that we did not expect values to be normally distributed here. A true normal curve has no x-axis limits. But recall that both predictivity and R^2 are constrained between hard limits of 0.0 – 1.0. This means normality should logically be impossible.

An arbitrary 100 models were generated to emulate each of the four Part I experimental model types, so 400 simulations were generated in total. This was about one-tenth as many runs as standard numerical method

dictates. It was enough to be reasonably stable, just not enough to be highly accurate. This limit was self-imposed, mainly because the simulation process could not be highly automated. Runs had to be done slowly, with SPSS syntax in batches (Appendix C). And, at this point, we were just trying to prove a point relative to the Part I research, rather than provide exhaustive results for an audience of professional statisticians.

Using the actual takeoff proportions from the Part I report, the following conditions were tested:

1. Low Financial Incentive	(Takeoff proportion = .300)	2 predictors, 1 constant
2. Low Financial Incentive	(Takeoff proportion = .300)	3 predictors, no constant
3. High Financial Incentive	(Takeoff proportion = .533)	2 predictors, 1 constant
4. High Financial Incentive	(Takeoff proportion = .533)	3 predictors, no constant

Notice that the constant was counted as one predictor here. All models were based on 30 cases (pilots), each with 60 available pseudo-predictor scores. Each such score was a normal (pseudo-) random number with mean of 5 and SD of 1. A "Takeoff" was coded as "1," a "Non-takeoff" as "0." Forward stepwise likelihood ratio (LR) was used as the predictor selection method, with the predictor inclusion criterion (PIN) set at .15 and the exclusion criterion (POUT) set at .20. These were simply the SPSS default values + .10, to be more liberal about allowing predictors into the model.[6] We had to do this because the Low Incentive base takeoff proportion was so high to start with (.70) that we knew that, otherwise, models with a constant would rely too heavily on that base rate, and models without a constant would fail to bootstrap.

RESULTS

Simulation Results

(Results are summarized in the abbreviated Table 2.) To illustrate the method here, the Low Financial Incentive model with just two predictors plus a constant (columns 2-3), had an average predictivity (μ_p) of .83. This meant that, on the average, logistic regression *on random numbers* successfully predicted 83% of takeoffs and accounted for 53% of the explainable (Nagelkerke) variance in the data. Note that in all cases, a standard rule of thumb was observed, namely that the number of model predictors k should not exceed $n/10$, 30/10 = 3. So this was exploring antecedent overfitting, not postcedent.

Average model performance was lower for the High Financial Incentive group. Moreover, models without a constant failed to converge in the High group. The immediate reason for this was bootstrap failure. Unless a model saw at least one predictor with probability of model improvement less than the PIN, it terminated before even getting started. No predictors were ever entered, and the model halted on the very first step.

The deeper reason for this bootstrap failure undoubtedly had to do with the High Incentive group's higher success ratio of takeoffs (.533). Theoretically, the hardest thing for a random number-based model to do should be to predict a perfectly random takeoff (.500 chance). We can visualize the underlying logical process by imagining one of its two hypothetical opposites—the case where takeoffs were .000. In that case, any model with a constant could predict takeoff perfectly by always guessing "No takeoff." The constant embodies a posteriori knowledge of the base rate of takeoff proportion, which captures the degree of uncertainty present in the dependent variable (DV). This uncertainty is greatest when takeoffs are 50% by chance and least when they are either 0 or 100%. Another way to view it is that, when *information* is defined as a condition of high certainty, there is literally *more information* in a success ratio of .300 than there is in one of .533 because .300 is closer to .000 (pure certainty). The logistic regression prediction equation takes advantage of this greater information, leading to better prediction from random sets with DV proportions either close to 0 or 1.

We could have avoided bootstrap failure, had we set the PIN sufficiently high. Every model would have then found some initial predictor to work with, no matter how poor, and the selection process could have continued. However, from experience, we knew that extremely high PINs (>.40) were often necessary to guarantee that all 100 models would bootstrap. This would have been absurdly relaxed in our entry criterion, so it was more apt to categorize these models as failures.

Finally, as a brief note on the distributions of μ and σ, as measured by standard skew and kurtosis (Fisher, 1970, ch. 3), normality was predictably unsupported. Appendix A graphically shows this.

Lacking a better method, confidence intervals (CI) for predictivity and R^2 means (μ) were roughly estimated by two methods. First, the usual z-score procedure of μ / SE_{μ}, the mean in question divided by its standard error, yielded one estimate. Second, it was possible to graph the values as a scatterplot and estimate the CI by visual inspection. Actually, these two estimates proved similar, given the models we examined (see Figure B1 in Appendix B).

So what did the confidence interval mean in this instance? Here, we wanted a .95 CI to mean that, if predictivity and R^2 exceeded the proper value, then there would be less than a 5% chance of this happening by accident on any given occasion. For a crude approximation, this method was adequate, provided we remain clear about its limitations.

DISCUSSION

There are really two main points to this exercise. First, the potential problem of antecedent overfitting is a serious issue, yet one far from common knowledge in many fields. It needs to be as understood and emphasized in applied experimental psychology as it is in fields such as economics. We do a large number of regression studies, and these ought to be as statistically sound as those of our best-versed colleagues.

Second, it was essential to defend the results from Part I of this study. We began with a large number of candidate predictors because we naturally wanted to examine many personality and demographic factors, looking for things that might illuminate pilot decision-making in the face of adverse weather. But, once we realized we had a methodological problem, it became a question of seeing it to conclusion. We had seen similar studies fail to recognize this issue; therefore, it made sense to bring it to light.

The first point—that a problem exists—has been amply addressed by example. To address the second point, consider the final best-model results first calculated for actual pilots in Part I.[7]

Now compare those predictivities and R^2s to those estimated by Monte Carlo simulation for random-data models. These had the exact same structure as the Part I models (same number of cases (n), candidate predictors (p), model predictors (k), and the same success ratio, that is, takeoffs/[takeoffs+non-takeoffs]). Table 4 summarizes.

Here, simulations were run for only two groups split by financial incentive, $n_{low}=28$ and $n_{high}=30$ groups. The truth is that combined-incentive models were not terribly illuminating because the Low and High Incentive subgroup best models seemed so different from one another. Financial incentive simply appeared to have too much effect on takeoff to make a combined-incentive model particularly meaningful.

It now became easy to see that the real-pilots Part I best-model results for High Financial Incentive pilots in no way exceeded what one would rightly expect by pure chance. This was in spite of its superficially significant Wald p value of .04. Real-pilots predictivity was about 75%—*lower* than average Monte Carlo predictivity with random numbers (76.3%). This was the reason for treating those results very gingerly in the Part I report.

So what about the Low Financial Incentive group best model, *Visibility x Ceiling + Constant*? What did it imply? Well, the answer to this should come in two parts. First—and unequivocally—weather does modulate takeoff rate. Pilots tend not to fly in bad weather, and that average effect was exactly what the constant was reflecting—the base rate. Seventy five percent of 28 pilots chose not to take off, whereas *every* pilot would certainly have taken off, given perfect weather and no other reason not to. Assuming a highly conservative base rate of 26/28 takeoffs for perfect weather, the estimated chance of getting the real takeoffs actually observed would be $p=1184040 \bullet .07^{21} \bullet .93^{7}$ by expansion of the binomial. That would be about four in ten billion billion. This was the exact reason a perfect-weather group was not tested in the first place. Why waste resources testing the obvious?

What the *VxC* part of the model was actually testing was fine weather discrimination. This involved variance left over after the base rate was taken into account. *VxC* was simply representing explainable variance unattributable to average weather foulness.

From a modeling perspective, the low incentive results implied that weather quality was primarily being perceived as a Go/No-go binary, threshold type of decision. The base rate (constant) supported that conclusion. To a lesser degree, some pilots seemed to think of weather as a continuum, probably a synergistic reaction between visibility and cloud ceiling. The *VxC* component supported that. To put it another way, their "cognitive whole" seemed greater than just the weighted sum of its individual parts. In pseudo-math, $\beta_{vc} V \times C > \beta_v V + \beta_c C$.

How reliable were these low incentive conclusions? Table 4 shows that the Part I real-pilots low-incentive 85.7% predictivity did exceed the random-generated Monte Carlo mean of 80.4%, although it did not top the estimate of 89% for the .95 CI. The real-pilots Nagelkerke R^2 of .52 considerably bested the Monte Carlo mean of .36, and came close to meeting the .95 CI of .59. So, judging from the Monte Carlo scatterplots (Appendix B, Figure B2), reliability for the low incentive n=28 experimental data was roughly α =.16 for predictivity and α =.08 for R^2.

Given that this was a preliminary study, one is free to draw one's own conclusions about the true reliability of the low-incentive *VxC* model. But do keep in mind that it does have clear face validity, being motivated by theory, not just by culling results from stepwise regression.

No matter what we decide about the *VxC* factor by itself, the two components of this model are important when considered together. The idea of a rule-based, threshold, cognitive process versus a synergistic, fine-discrimination process is certainly a useful heuristic to guide future research in decision making. It would apply broadly to all kinds of decision making, not just aviation weather research.

Now, finally, what to say about the influence of money? The *absence* of effects for the High Financial Incentive group was, oddly enough, an interesting result. More precisely, the base rate of 46.7% non-takeoffs (100–53.3) did imply a strong *average* weather effect (expected

p=145422675•.07^{14}•.93^{16}≅ 3•10^{-9}). However, assuming reliable low-incentive fine *VxC* weather discrimination, then the inability to find the same fine discrimination in the high incentive condition implied that the financial incentive completely destroyed this. This absence of fine discrimination was important. It implied that, as soon as money entered the picture, all distinction between various degrees of bad weather ceased. In plain language, money probably disables fine discrimination, at least as far as weather goes. These results may generalize to many other domains as well. We certainly know, anecdotally, that people do all kinds of foolish things for money. Here we see just one example of that general principle.

To wrap this up, the Part I results were not fatally flawed by the large number of candidate predictors. However, the problem of antecedent overfitting did need to be factored in. Once it was, then we had a much more honest picture of what was likely to be reliable.

CONCLUSIONS

Overfitting is a common problem in regression studies. During the course of our weather research, we discovered that there are at least two major kinds of overfitting,[8] which we subsequently chose to call antecedent and postcedent overfitting. "Antecedent" refers here to the situation existing prior to data analysis, after candidate predictors have been measured, but before modeling starts. "Postcedent" refers to the situation after modeling concludes. Postcedent overfitting, therefore, refers to the situation where too many predictors were included in a given regression model. Antecedent overfitting refers to the situation where too many candidate predictors were present before modeling began.

Postcedent overfitting is universally known. Antecedent overfitting is known to statistical theorists and in a few domains such as economics but is quite new to us in the social sciences. Hopefully, the remainder of the research community will follow the lead of Foster, Stine, and others in treating this as a serious issue. Antecedent overfitting was encountered here by accident and would have compromised the Part I study, had it not been recognized and confronted. Using a large number of candidate predictors does not have to be a fatal error. It just needs to be treated knowingly as part and parcel of the design and analysis in question.

From a practitioner viewpoint, there are basically two ways to handle antecedent overfitting. First, we can minimize the problem ahead of time by limiting the number of candidate predictors we measure. Second, we can deal with it post hoc, by running custom Monte Carlo simulations set up with the same number of cases (n), candidate predictors (p), model predictors (k), and success ratios (S) as the experimental data. The predictivity and R^2 mean scatterplots of these custom simulations will allow a rough estimate of .95 confidence intervals, against which we can compare the actual predictivity and R^2 of our real-data models. This is essentially an ad hoc way of doing what Rencher and Pun (1980) did in closed form for standard least-squares regression.

The admitted problem with trying to limit p is that there is not yet a truly simple, reliable rule of thumb to do it, and to create one is beyond the scope of this paper and the mission of this research. What we are talking about is finding mathematical functions of the form $\mu_{.95} = f(p, k, n, S)$ that could accept four numbers as input, and then tell us the values for predictivity and R^2 we would have to exceed to get 95% reliability in spite of p. This is a 4–dimensional function and would require hundreds of thousands of Monte Carlo simulations to cover a full range of values for all four dimensions.

The problem with the post hoc approach is that it means running the experiment and *then* worrying about whether the results are reliable or not. What do we do if we find "significant" predictors that later totally fail stricter Monte Carlo-based significance tests? As we saw, this was not hard to do, particularly when predictivities in the 80-90% range could be the result of random numbers.

In the end, this process of estimating p is obviously a tradeoff, but one we are uncertain about at this point in time. The short answer is that there probably are "sweet spots" representing sufficient predictors to be useful without sacrificing too much in the way of reliability. We simply do not know what those numbers are yet. In the meantime, the rule of thumb can only be something like "Use the smallest predictor set possible, probably ten or less." Failing this, if many predictors are intentionally used on a first pass, then it should be followed up with a confirmatory study testing the ten or so strongest. Anything that survives both studies is likely to be authentic.

REFERENCES

Derksen, S. & Keselman, H.J. (1992). Backward, forward, and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45*, 265-82.

Fisher, R.A. (1970). *Statistical Methods for Research Workers* (14th ed.). Edinburgh; Tweeddale Court.

Foster, D.P., & Stine, R.A, (1998). Honest confidence intervals for the error variance in stepwise regression. Retrieved July 23, 2003, from http://www-stat.wharton.upenn.edu/~bob/research/honests2.pdf . Unpublished manuscript, University of Pennsylvania, Wharton School, Department of Statistics.

Kendall, M.G., & Stuart, A. (1961). *The advanced theory of statistics*, Vol. 2. New York: Hafner.

Kreyszig, E. (1972). *Advanced engineering mathematics.* (3rd Ed.). New York: Wiley.

Microsoft. (1999). *Excel 2000*. Seattle: Microsoft Corp.

Norušis, M.J. (1999). *SPSS regression models 10.0*. Chicago: SPSS, Inc.

Press, W.H., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T. (1988). *Numerical recipes in C.* New York: Cambridge.

Rencher, A.C., & Pun, F.C. (1980). Inflation of R^2 in best subset regression. *Technometrics, 22*(1), 49-53.

SPSS, Inc. (2003). Logistic regression. Operational algorithms retrieved February 17, 2004 from http://support.spss.com/tech/default.asp. Unpublished manuscript, Chicago: SPSS, Inc.[9]

Tabachnick, B.G., Fidell, L.S. (2000). *Multivariate statistics* (4th Ed.). Needham Heights, MA: Allyn & Bacon.

ENDNOTES

[1] In this report, the word "predictivity" is used as a proxy for the formal statistical terms sensitivity and specificity (R.A. Stine, personal communication, March 16, 2004). Sensitivity here refers to the number of correctly predicted takeoffs. Specificity is the number of correctly predicted non-takeoffs. In signal detection theory, these would be Hits and Correct Rejections, respectively. So predictivity = (sensitivity + specificity)/(total cases). We use predictivity primarily to simplify description of overall model performance by using a single term to describe "total takeoffs and non-takeoffs that a given model correctly predicted."

[2] R.A. Stine (personal communication, March 16, 2004) points out that the order of entering model predictors does not matter; therefore there are only 1/6 as many models as there otherwise would be.

[3] Technically, this depends on how the model is set up. For instance, it can be set up to hill-climb in likelihood ratio space (or hill-descend in Wald p space). But, for ease of understanding, it is easier to talk about hill-climbing in predictivity space, and it is nearly as accurate.

[4] Because it is based on likelihood ratio estimates, and not strictly on sums of squares, not all statisticians agree that R^2 is as meaningful in logistic regression as it is in standard regression.

[5] Bootstrapping has various and special meanings within statistics. Here, we are merely using it in the sense of "hauling yourself up by your own bootstraps," that is, to get some process off and running.

[6] R.A. Stine (personal communication, March 16, 2004) points out that this PIN was "...essentially...the AIC criterion (Akaike Information Criterion)...AIC has problems with overfitting in [the] context of a 'wide' data set (one with as many or more columns as rows)." Here, our rows were pilots (n=30) and columns were the pseudo-predictors (n=60).

[7] The final Low Financial Incentive group data represent two outliers being dropped because the pilots had made statements implying they had not taken the study seriously, so n was reduced from 30 to 28.

[8] There is also a third, which might be called "manifold overfitting." This involves the issue of adjusting model significance based on the number of models explored. The more models we test, the more likely some are to be "significant" by chance. Stine (personal communication, January 26, 2004) suggests a Bonferroni-type correction for this. To oversimplify, Bonferroni approaches adjust the critical significance (e.g. =.05) by dividing it by the number of elements tested (for our purposes, the number of models explored). Needless to say, Bonferroni corrections favor modeling based on theory, and greatly penalize "shotgun" approaches where many models are examined with no underlying theory at all.

[9] Note: You must be a registered SPSS user to access the SPSS technical support site.

FIGURES AND TABLES

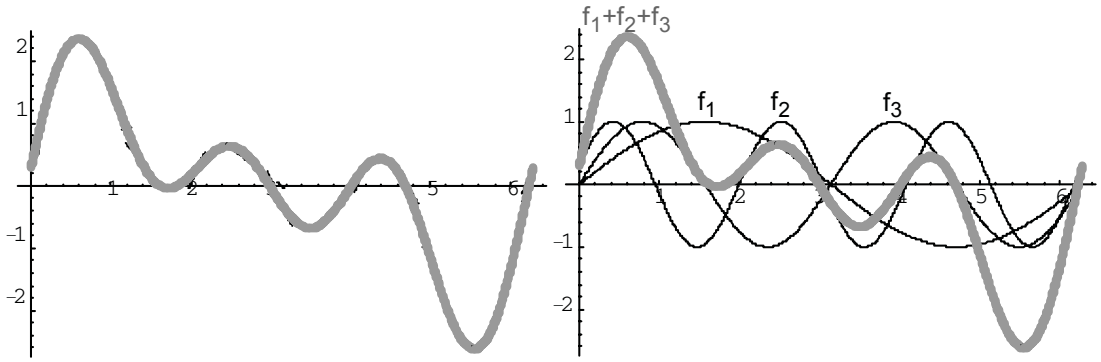

Figure 1. A signal that superficially looks very complex can actually be broken down into three simple components, f_1, f_2, and f_3.

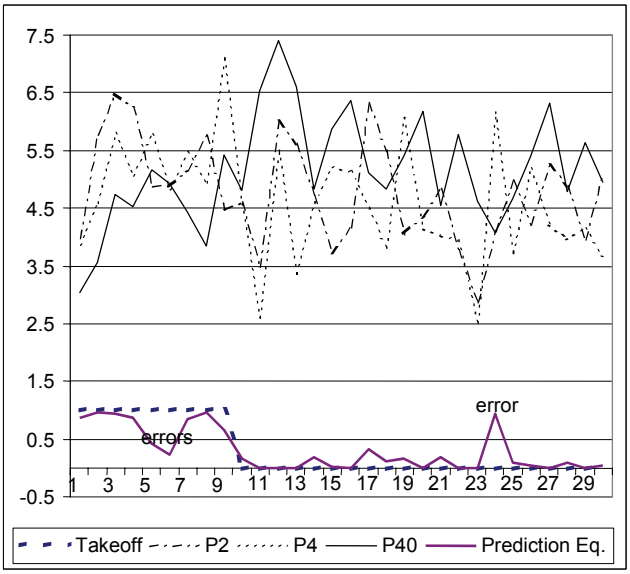

Figure 2. A graph of the three best pseudo-predictor scores. The y-axis represents score values. Each of the 30 pilots listed on the x-axis has three random-number score values on the y-axis. These random scores, entered into a regression model, seemed capable of predicting takeoff at better than chance level.

Table 1. An abbreviated view of sample simulated predictor scores for *n*=30 "pilots," each of which "took off" (1) or did not (0), and also had 60 simulated predictor scores.

Pilot	Takeoff	Predictor #				
(case) #	0=No, 1=Yes	1	2	3	...	60
1	0	3.72	5.24	6.28	...	6.73
2	0	4.77	6.10	3.91	...	3.31
...	...	...	...	...	...	...
30	1	4.63	4.67	4.63		4.91

Table 2. Summary statistics for Monte Carlo simulation of logistic regression modeling. The *N*=60 data were divided into two *n*=30 groups by *Financial Incentive*. The most important result is that average predictivities (μ, in gray) all exceeded the chance level of .50, even though the input predictor scores were essentially random numbers. This is an unwanted artifact of stepwise regression.

	Low Financial Incentive group				High Financial Incentive group			
	Success ratio of takeoffs = 0.300				Success ratio of takeoffs = 0.533			
DATA	models with constant		without constant		models with constant		without constant	
Run #	correct preds	Nagel R^2	correct preds	Nagel R^2	correct preds	Nagel R^2	correct preds	Nagel R^2
1	0.767	0.529	0.900	0.822	0.733	0.369	failed	
…	…	…	…	…	…	…	…	
100	0.867	0.508	0.700	0.575	0.733	0.474	failed	
				RESULTS				
μ	*0.83*	*0.53*	*0.84*	*0.67*	*0.76*	*0.48*		
σ	0.05	0.12	0.05	0.09	0.05	0.10		
skew	-0.27	0.19	-0.47	0.09	0.39	0.98		
SE_{skew}	0.24	0.24	0.24	0.24	0.24	0.24		
p_{skew}	0.13	0.22	0.027	0.35	0.051	0.000		
kurt	0.07	-0.50	0.44	0.26	0.21	2.15		
SE_{kurt}	0.48	0.48	0.48	0.48	0.48	0.48		
p_{kurt}	0.44	0.15	0.18	0.29	0.33	0.000		
CI 95	≅.92	≅.72	≅.92	≅.82	≅.85	≅.64		

Table 3. Results from the Part I study, showing best models for the Low and High Financial Incentive groups. "Best" was defined as a combination of low Wald *p*-value, high predictivity, and model simplicity. Experience shows that models lacking all three qualities often fail to perform reliably on new data.

Data set	$p_{takeoffs}$	Best model found	Wald *p*	Predictivity
Low $ Incentive		*Visibility x Ceiling*	.008	
N=28	0.250	*Constant*	.003	85.7%
High $ Incentive		*Financial Motivation (buck_mot)*		
N=30	0.533	*x Predictor P*	≈ .04	≈ 75%
		Constant		

Table 4. If we run the same models with random numbers many times, the average (μ) predictivities, R^2s, and upper confidence intervals (CI .95) give us baselines against which to compare the reliability of models based on actual human data.

	Low Fin. Incentive group		High Fin. Incentive group	
	models with constant		models with constant	
	Predictivity	Nagel R^2	Predictivity	Nagel R^2
$\mu_{MonteCarlo}$	80.4	0.36	76.3	0.48
CI 95	≅.89	≅.59	≅.85	≅.64
$\mu_{ActualData}$	85.7	0.52	75	0.28
α estimated	0.16	0.08	NS	NS

APPENDIX A

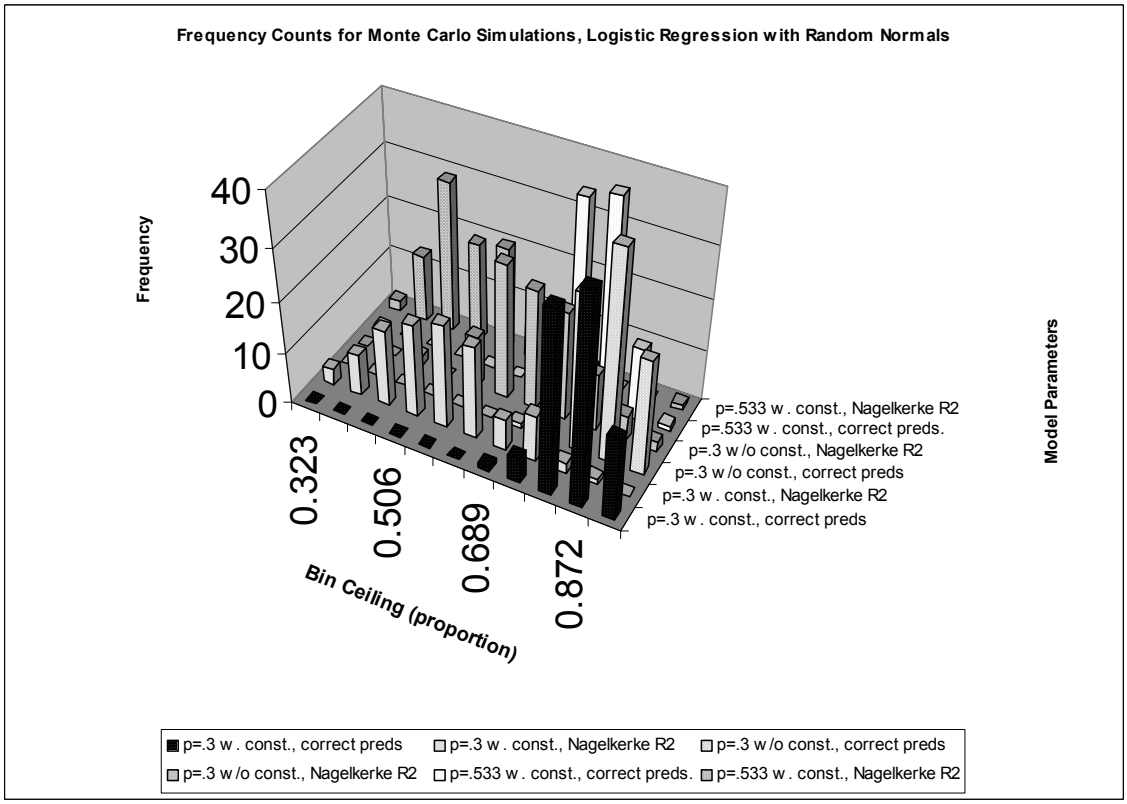

Figure A1. Frequency counts for Monte Carlo simulations, SPSS forward stepwise logistic regression, 30 cases, 60 predictors. Because predictivity and R^2 are range-limited, 0-1, we do not expect these distributions to be normal. They behave more like beta functions, the distributions often being bunched up to either the right or the left. Here we can visually see this happening.

APPENDIX B

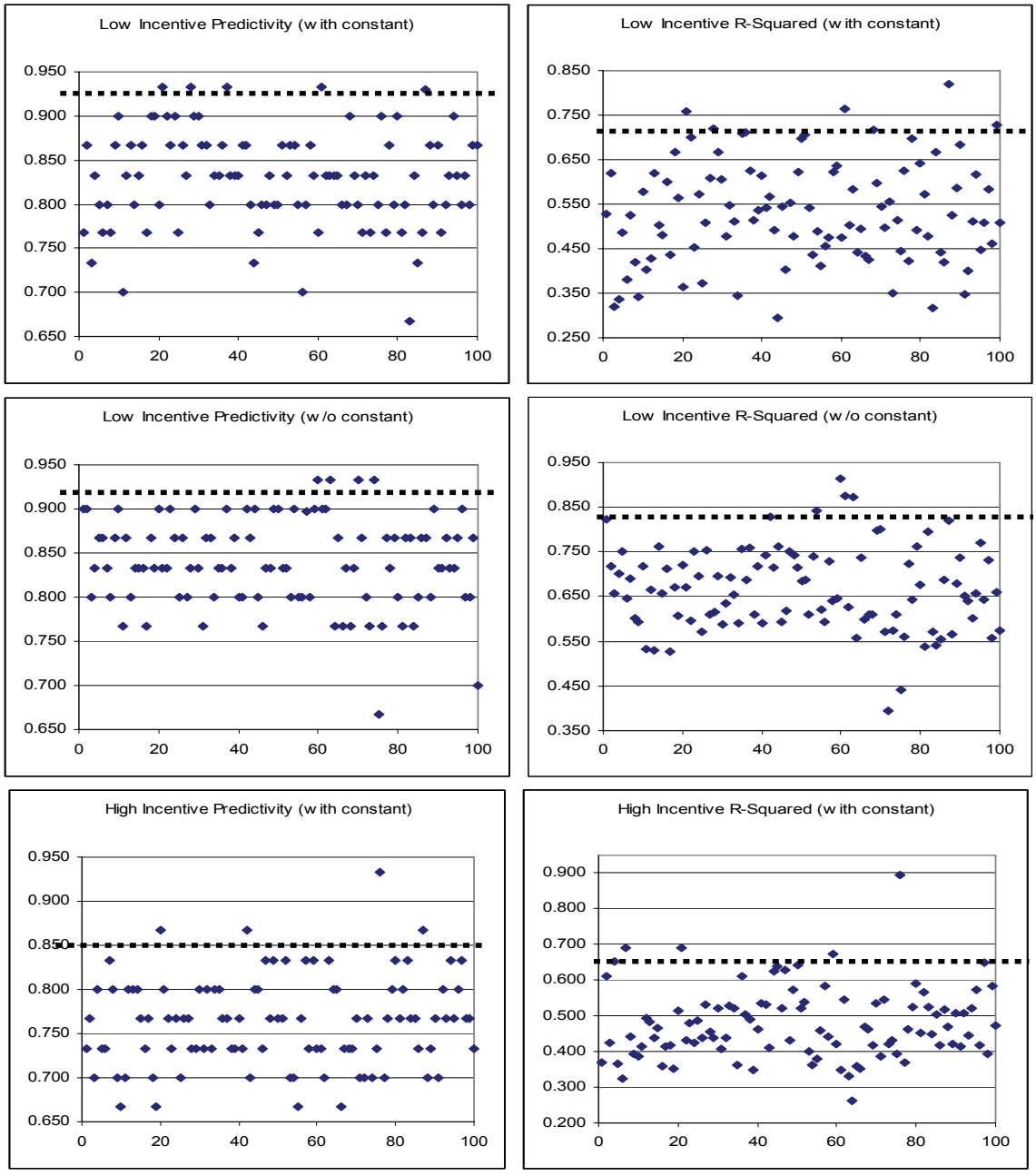

Figure B1. Scatterplots of predictivity and Nagelkerke R^2 values obtained during the n=30 Monte Carlo simulations (before outliers were eliminated in the Low Financial Incentive group). The x-axis is the simulation number, 1 being the first simulation and 100 the last, for that combination of conditions. The y-axis is the corresponding value of predictivity or R^2 obtained when running each model with random numbers. These scatterplots allow us to estimate the 95% confidence interval (.95 CI) by inspection (the dashed line on each plot). We can then test an empirically obtained value of predictivity or R^2 by seeing if it meets or exceeds the appropriate .95 CI value.

Table B1. The scatterplot estimates of Figure 4, arranged in table form. Table 2 contains the same data.

	Low Financial Incentive group					High Financial Incentive group			
	models with constant		without constant			models with constant		without constant	
	correct preds	Nagel R^2	correct preds	Nagel R^2		correct preds	Nagel R^2	correct preds	Nagel R^2
CI 95	≅.92	≅.72	≅.92	≅.82		≅.85	≅.64	failed	

Finally, we present one last set of Monte Carlo estimates. We examined one n=28 model, having determined the need to eliminate two outliers in the Low Financial Incentive group. This led to a model *Visibility x Ceiling + constant*, with takeoff proportion = .25, and the following scatterplots:

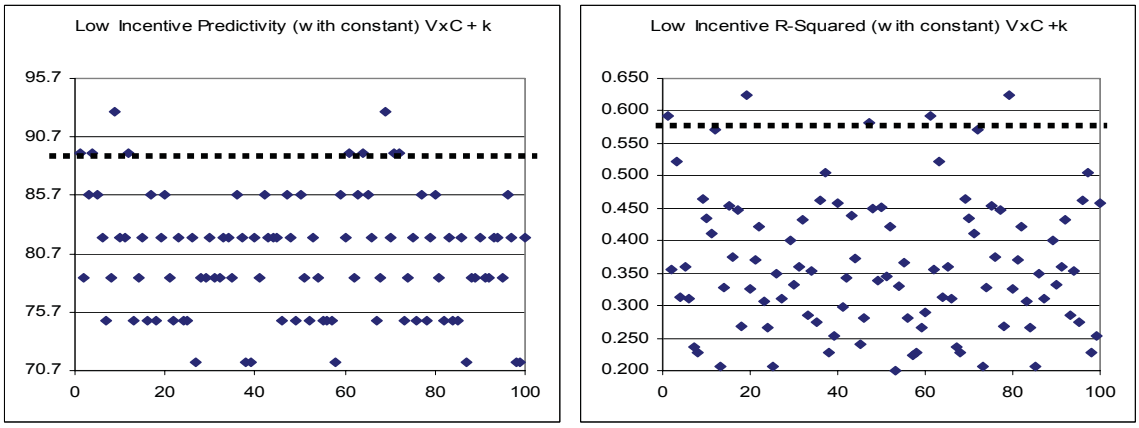

Figure B2. Scatterplots used in the Part I study to re-estimate predictivity and R^2 after the elimination of two outliers in the Low Financial Incentive group.

APPENDIX C

Below is an example of the SPSS syntax used to generate random numbers and run the logistic regression simulations. GET FILE only had to be called once. Otherwise, the rest of the commands were repeated, to execute in batches of ten simulations per run. Here, the syntax is arranged in two columns, to fit on a single page. To actually run this syntax, it needs to be arranged in one, continuous column.

```
GET FILE = 'c:\Billy Bob\Math &
Statistics\Logistic Regression\zoot.sav'.
COMPUTE p1=RV.NORMAL(5,1).
COMPUTE p2=RV.NORMAL(5,1).
COMPUTE p3=RV.NORMAL(5,1).
COMPUTE p4=RV.NORMAL(5,1).
COMPUTE p5=RV.NORMAL(5,1).
COMPUTE p6=RV.NORMAL(5,1).
COMPUTE p7=RV.NORMAL(5,1).
COMPUTE p8=RV.NORMAL(5,1).
COMPUTE p9=RV.NORMAL(5,1).
COMPUTE p10=RV.NORMAL(5,1).
COMPUTE p11=RV.NORMAL(5,1).
COMPUTE p12=RV.NORMAL(5,1).
COMPUTE p13=RV.NORMAL(5,1).
COMPUTE p14=RV.NORMAL(5,1).
COMPUTE p15=RV.NORMAL(5,1).
COMPUTE p16=RV.NORMAL(5,1).
COMPUTE p17=RV.NORMAL(5,1).
COMPUTE p18=RV.NORMAL(5,1).
COMPUTE p19=RV.NORMAL(5,1).
COMPUTE p20=RV.NORMAL(5,1).
COMPUTE p21=RV.NORMAL(5,1).
COMPUTE p22=RV.NORMAL(5,1).
COMPUTE p23=RV.NORMAL(5,1).
COMPUTE p24=RV.NORMAL(5,1).
COMPUTE p25=RV.NORMAL(5,1).
COMPUTE p26=RV.NORMAL(5,1).
COMPUTE p27=RV.NORMAL(5,1).
COMPUTE p28=RV.NORMAL(5,1).
COMPUTE p29=RV.NORMAL(5,1).
COMPUTE p30=RV.NORMAL(5,1).
COMPUTE p31=RV.NORMAL(5,1).
COMPUTE p32=RV.NORMAL(5,1).
COMPUTE p33=RV.NORMAL(5,1).
COMPUTE p34=RV.NORMAL(5,1).
COMPUTE p35=RV.NORMAL(5,1).
COMPUTE p36=RV.NORMAL(5,1).
COMPUTE p37=RV.NORMAL(5,1).
COMPUTE p38=RV.NORMAL(5,1).
COMPUTE p39=RV.NORMAL(5,1).
COMPUTE p40=RV.NORMAL(5,1).
COMPUTE p41=RV.NORMAL(5,1).
COMPUTE p42=RV.NORMAL(5,1).
COMPUTE p43=RV.NORMAL(5,1).
COMPUTE p44=RV.NORMAL(5,1).
COMPUTE p45=RV.NORMAL(5,1).
COMPUTE p46=RV.NORMAL(5,1).
COMPUTE p47=RV.NORMAL(5,1).
COMPUTE p48=RV.NORMAL(5,1).
COMPUTE p49=RV.NORMAL(5,1).
COMPUTE p50=RV.NORMAL(5,1).
COMPUTE p51=RV.NORMAL(5,1).
COMPUTE p52=RV.NORMAL(5,1).
COMPUTE p53=RV.NORMAL(5,1).
COMPUTE p54=RV.NORMAL(5,1).
COMPUTE p55=RV.NORMAL(5,1).
COMPUTE p56=RV.NORMAL(5,1).
COMPUTE p57=RV.NORMAL(5,1).
COMPUTE p58=RV.NORMAL(5,1).
COMPUTE p59=RV.NORMAL(5,1).
COMPUTE p60=RV.NORMAL(5,1).
LOGISTIC REGRESSION TAKEOFF WITH
p1 p2 p3 p4 p5 p6 p7 p8 p9 p10 p11 p12
p13 p14 p15 p16 p17 p18 p19 p20 p21 p22
p23 p24 p25 p26 p27 p28 p29 p30 p31 p32
p33 p34 p35 p36 p37 p38 p39 p40 p41 p42
p43 p44 p45 p46 p47 p48 p49 p50 p51 p52
p53 p54 p55 p56 p57 p58 p59 p60
/METHOD FSTEP(LR)
/CRITERIA   PIN(.15) POUT(.20) CUT(.5)
/PRINT SUMMARY.
```